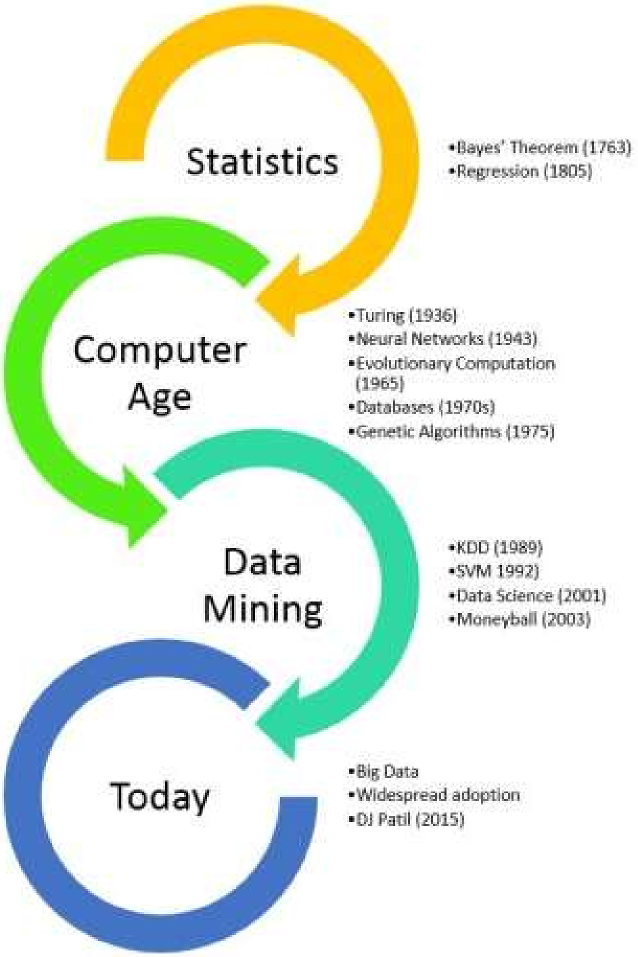
Data Preprocessing in Data Mining
Data Transformation

Dr. Tran Anh Tuan,

Faculty of Mathematics and Computer Science,

University of Science, HCMC



Statistics

- Bayes' Theorem (1763)
- Regression (1805)

Computer Age

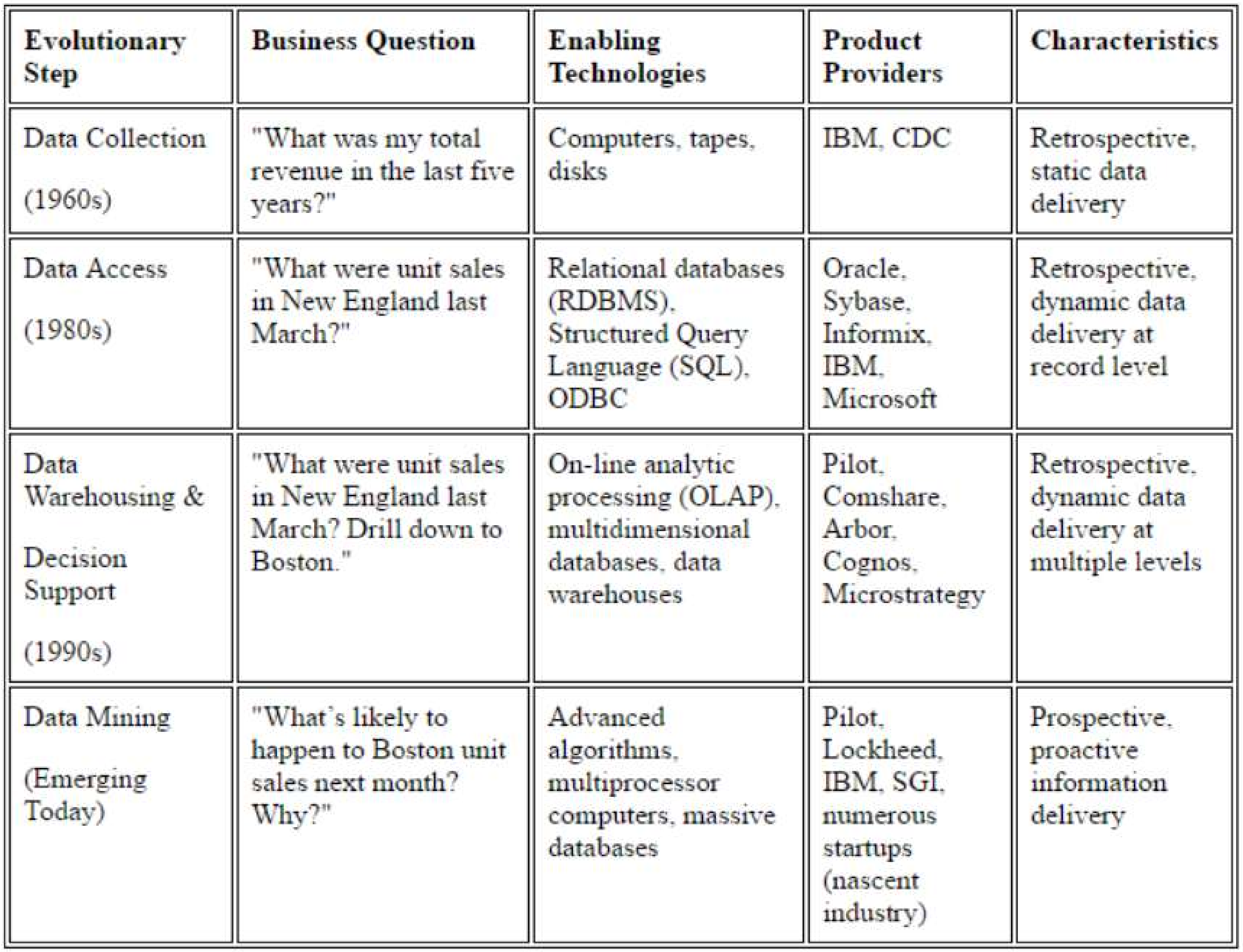
- Turing (1936)
- Neural Networks (1943)
- Evolutionary Computation (1965)
- Databases (1970s)
- Genetic Algorithms (1975)

Data Mining

- KDD (1989)
- SVM 1992)
- Data Science (2001)
- Moneyball (2003)

Today

- Big Data
- Widespread adoption
- DJ Patil (2015)

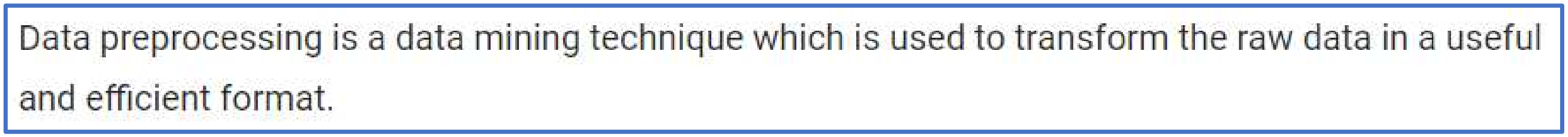


Evolutionary Step	Business Question	Enabling Technologies	Product Providers	Characteristics
Data Collection (1960s)	"What was my total revenue in the last five years?"	Computers, tapes, disks	IBM, CDC	Retrospective, static data delivery
Data Access (1980s)	"What were unit sales in New England last March?"	Relational databases (RDBMS), Structured Query Language (SQL), ODBC	Oracle, Sybase, Informix, IBM, Microsoft	Retrospective, dynamic data delivery at record level
Data Warehousing & Decision Support (1990s)	"What were unit sales in New England last March? Drill down to Boston."	On-line analytic processing (OLAP), multidimensional databases, data warehouses	Pilot, Comshare, Arbor, Cognos, Microstrategy	Retrospective, dynamic data delivery at multiple levels
Data Mining (Emerging Today)	"What's likely to happen to Boston unit sales next month? Why?"	Advanced algorithms, multiprocessor computers, massive databases	Pilot, Lockheed, IBM, SGI, numerous startups (nascent industry)	Prospective, proactive information delivery

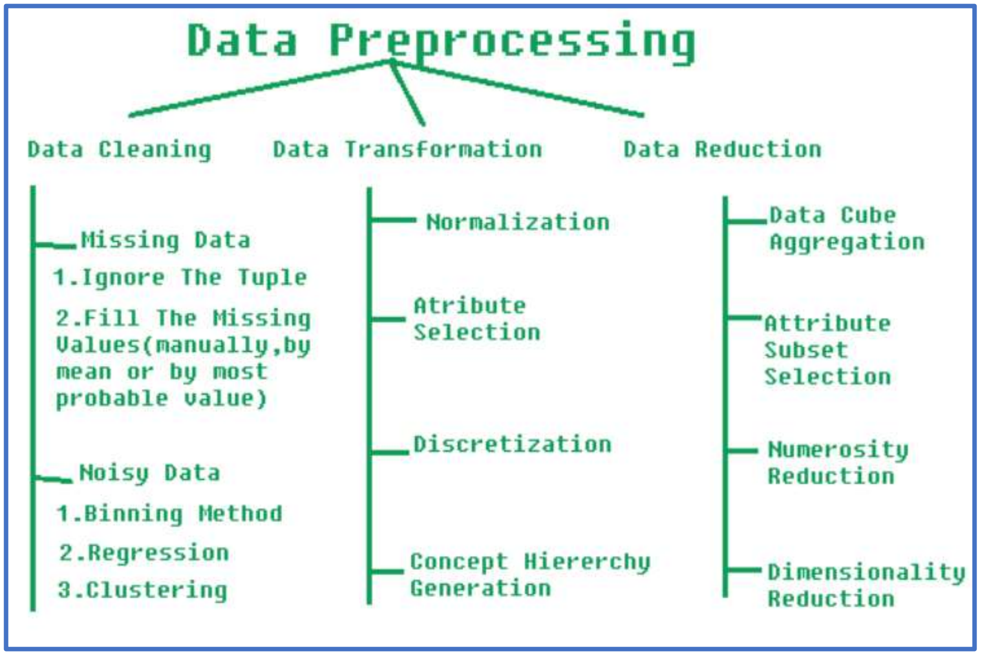
Syllabus

- **Lecture 1 : Data Preprocessing**
- Lecture 2 : Explanatory Data Analysis
- Lecture 3 : Feature Engineering (Feature Importance and Selection)
- Lecture 4 : Association Rule Learning
- Lecture 5 : Unsupervised Clustering
- Lecture 6 : Unsupervised Clustering (cont.)
- Lecture 7 : Anomaly and Outlier Detection
- Lecture 8 : Regression and Classification Learning
- Lecture 9 : Recommendation Learning
- Lecture 10 : Final Project Requirement

Introduction



Data preprocessing is a data mining technique which is used to transform the raw data in a useful and efficient format.



Data Preprocessing

Data Cleaning

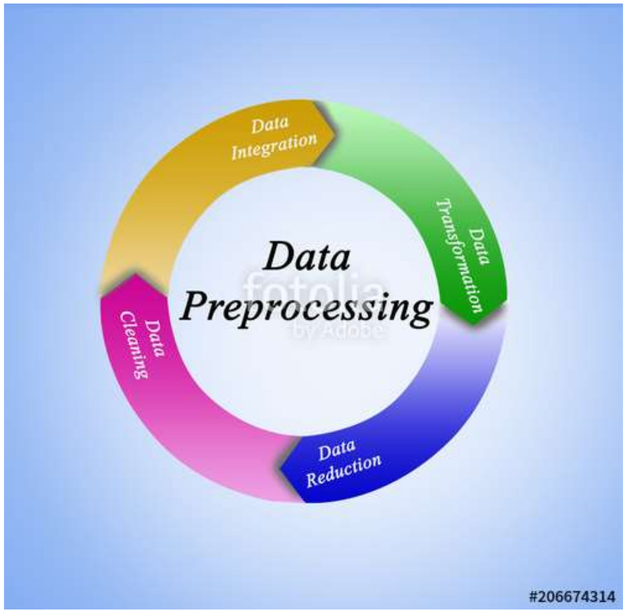
- Missing Data
  1. Ignore The Tuple
  2. Fill The Missing Values(manually,by mean or by most probable value)
- Noisy Data
  1. Binning Method
  2. Regression
  3. Clustering

Data Transformation

- Normalization
- Atribute Selection
- Discretization
- Concept Hierercy Generation

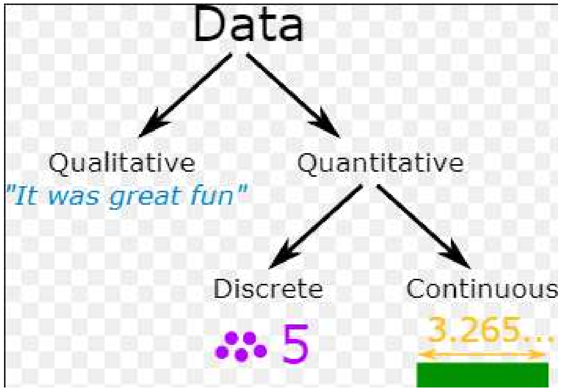
Data Reduction

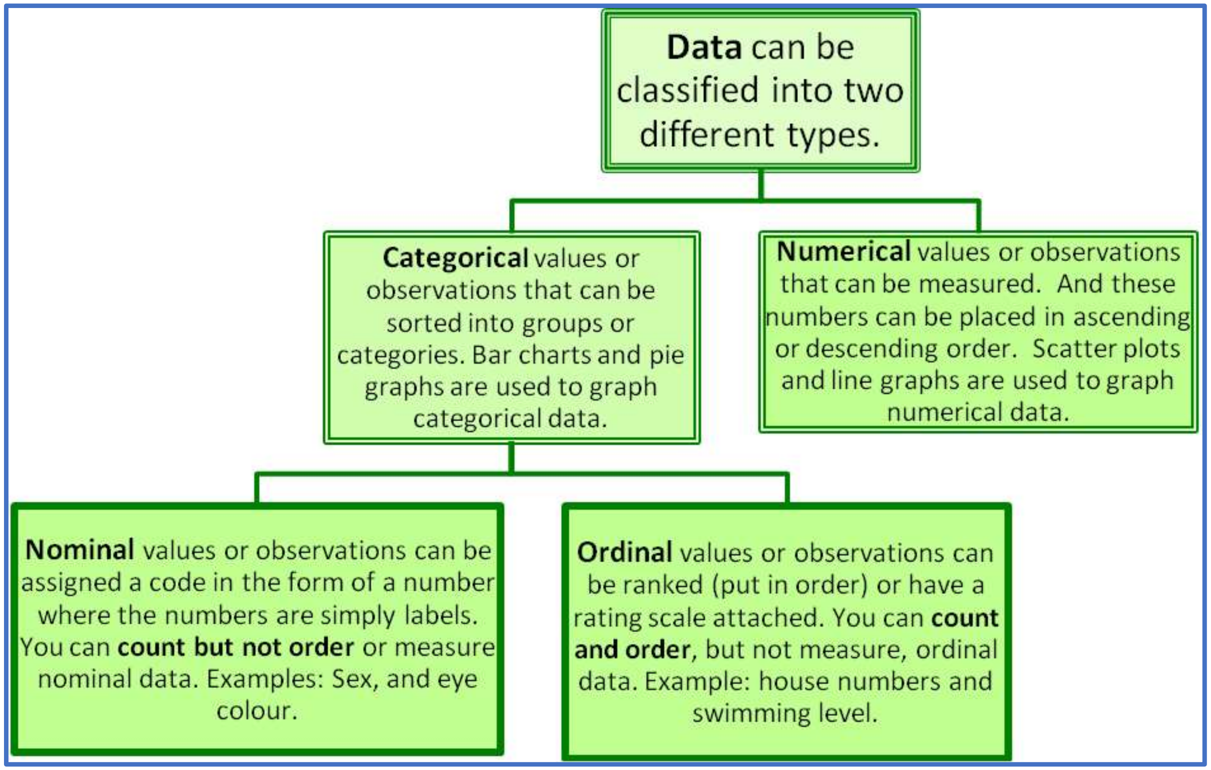
- Data Cube Aggregation
- Attribute Subset Selection
- Numerosity Reduction
- Dimensionality Reduction



Data Types

- When working with statistics, it's important to recognize the different types of data: numerical (discrete and continuous), categorical, and ordinal.





Data can be classified into two different types.

Categorical values or observations that can be sorted into groups or categories. Bar charts and pie graphs are used to graph categorical data.

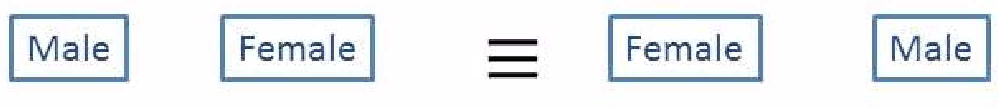
Numerical values or observations that can be measured. And these numbers can be placed in ascending or descending order. Scatter plots and line graphs are used to graph numerical data.

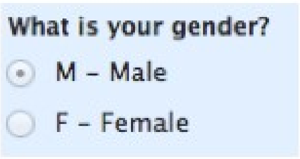
Nominal values or observations can be assigned a code in the form of a number where the numbers are simply labels. You can **count but not order** or measure nominal data. Examples: Sex, and eye colour.

Ordinal values or observations can be ranked (put in order) or have a rating scale attached. You can **count and order**, but not measure, ordinal data. Example: house numbers and swimming level.

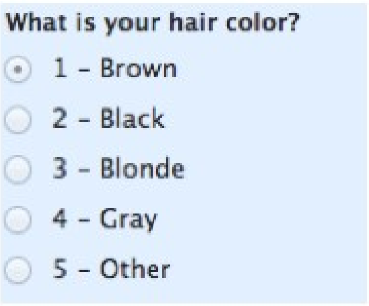
Nominal

- Values represent discrete units
- Changing the order of units does not change their value

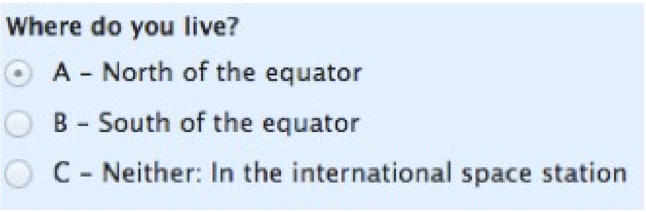



What is your gender?

● M – Male
○ F – Female


What is your hair color?

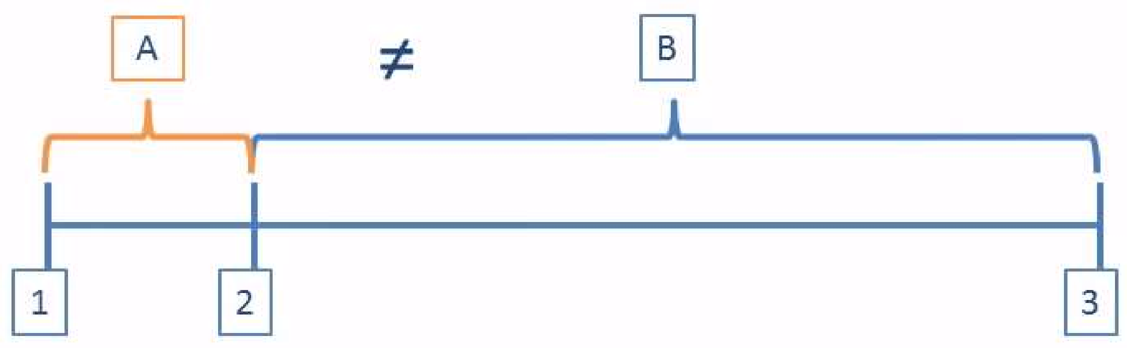
● 1 – Brown
○ 2 – Black
○ 3 – Blonde
○ 4 – Gray
○ 5 – Other

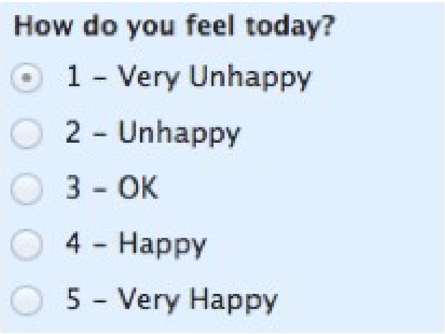

Where do you live?

● A – North of the equator
○ B – South of the equator
○ C – Neither: In the international space station

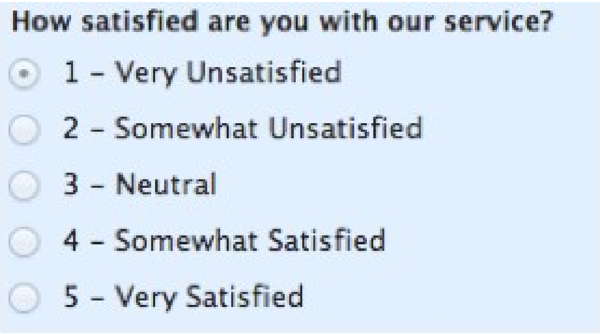
Ordinal

- Values represent discrete and ordered units
- Distance between units is not the same




How do you feel today?

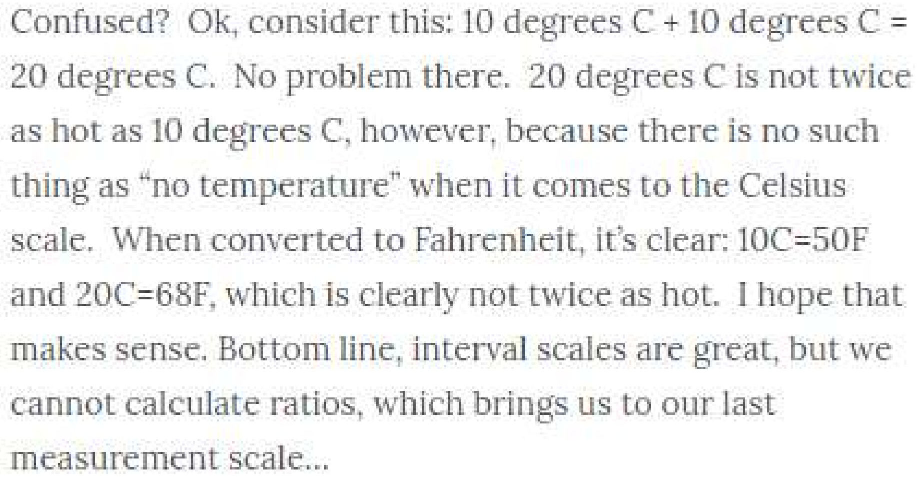
● 1 – Very Unhappy
○ 2 – Unhappy
○ 3 – OK
○ 4 – Happy
○ 5 – Very Happy


How satisfied are you with our service?

● 1 – Very Unsatisfied
○ 2 – Somewhat Unsatisfied
○ 3 – Neutral
○ 4 – Somewhat Satisfied
○ 5 – Very Satisfied

Interval

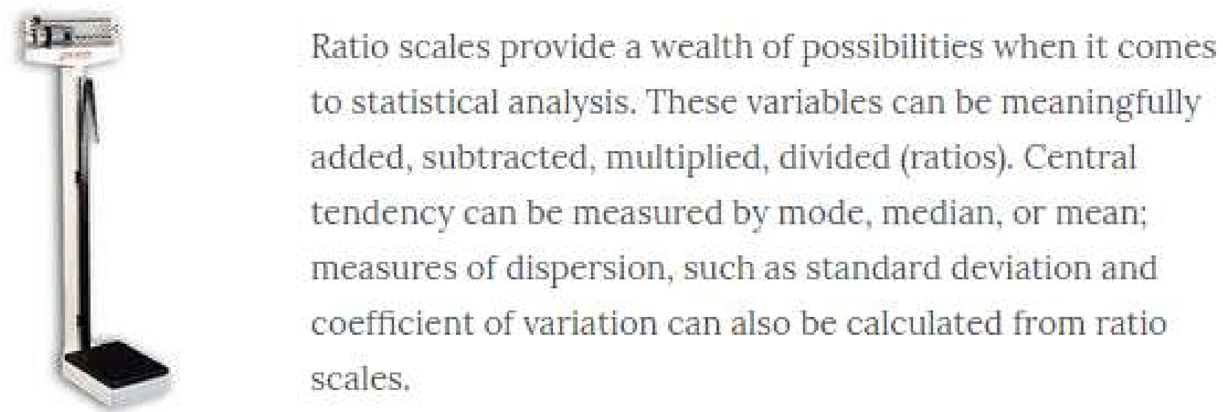
- Ordered units with intermediate values
- Distance between units is the same
- No absolute zero
  - Origin is arbitrary
  - A person with an IQ score of 160 is NOT twice as smart as someone with an IQ score of 80



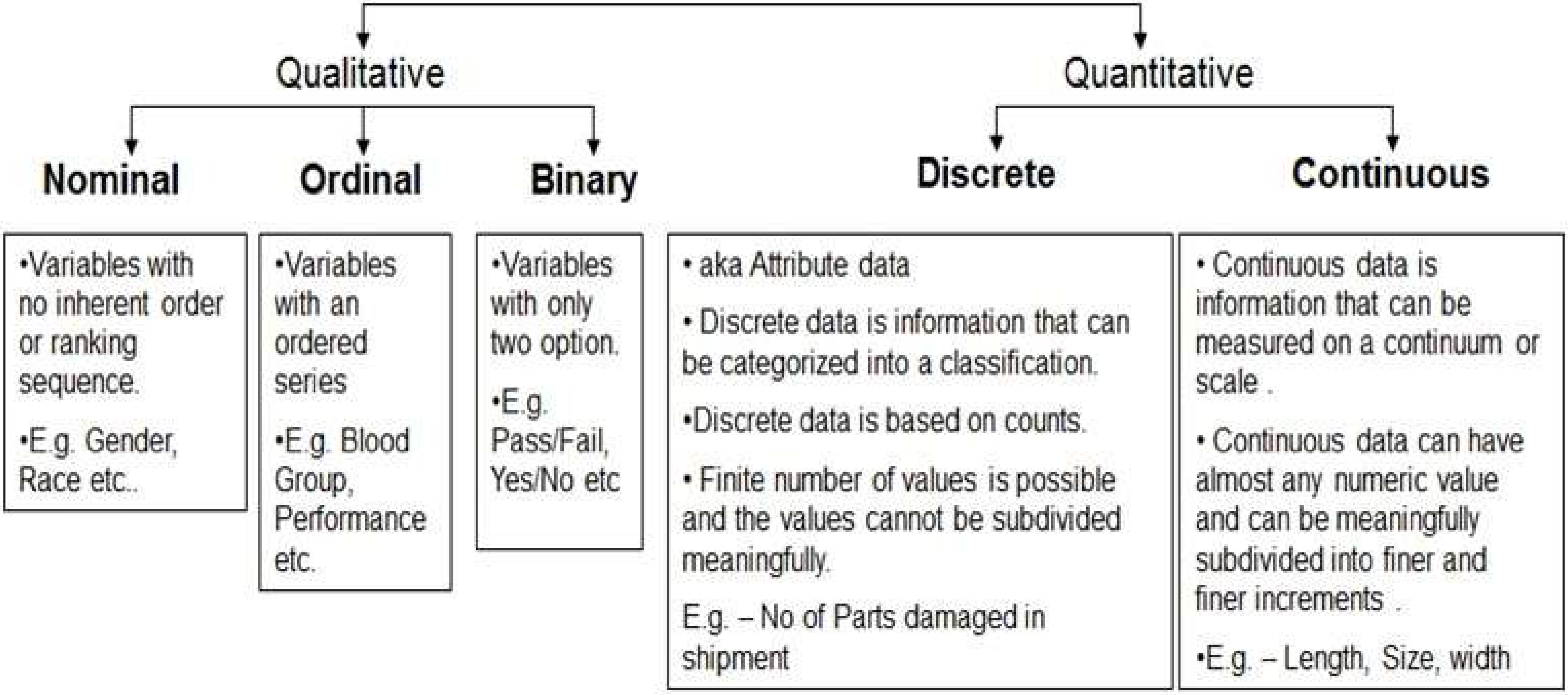
Confused? Ok, consider this: 10 degrees C + 10 degrees C = 20 degrees C. No problem there. 20 degrees C is not twice as hot as 10 degrees C, however, because there is no such thing as "no temperature" when it comes to the Celsius scale. When converted to Fahrenheit, it's clear: 10C=50F and 20C=68F, which is clearly not twice as hot. I hope that makes sense. Bottom line, interval scales are great, but we cannot calculate ratios, which brings us to our last measurement scale...

Ratio

- Ordered units with intermediate values
- Distance between units is the same
- Absolute zero
  - Origin is at zero
  - A 12 inch long sandwich is twice the length of a 6 inch sandwich. (But, is it twice as delicious ?)



Ratio scales provide a wealth of possibilities when it comes to statistical analysis. These variables can be meaningfully added, subtracted, multiplied, divided (ratios). Central tendency can be measured by mode, median, or mean; measures of dispersion, such as standard deviation and coefficient of variation can also be calculated from ratio scales.



Qualitative

Quantitative

Nominal

- Variables with no inherent order or ranking sequence.
- E.g. Gender, Race etc..

Ordinal

- Variables with an ordered series
- E.g. Blood Group, Performance etc.

Binary

- Variables with only two option.
- E.g. Pass/Fail, Yes/No etc

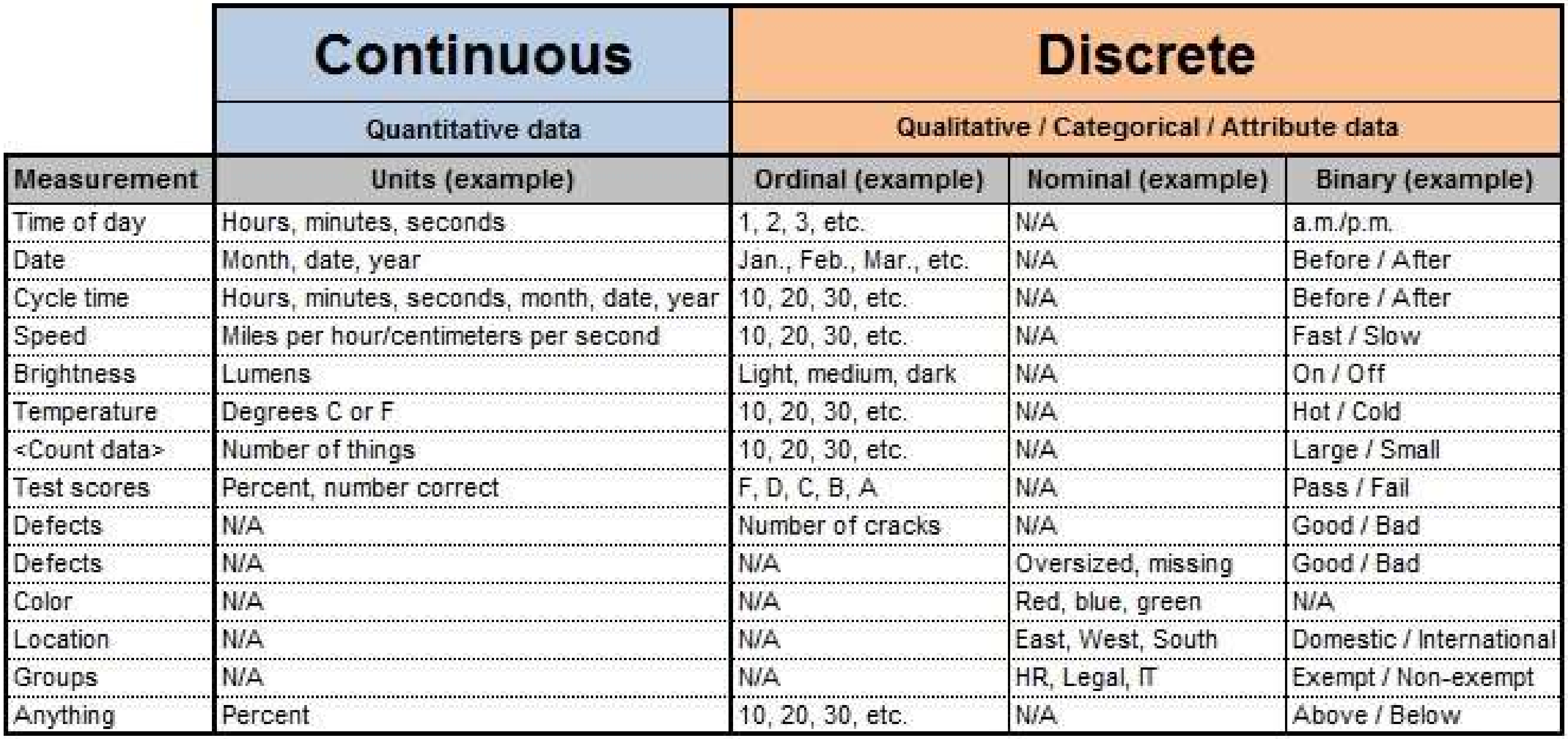
Discrete

- aka Attribute data
- Discrete data is information that can be categorized into a classification.
- Discrete data is based on counts.
- Finite number of values is possible and the values cannot be subdivided meaningfully.

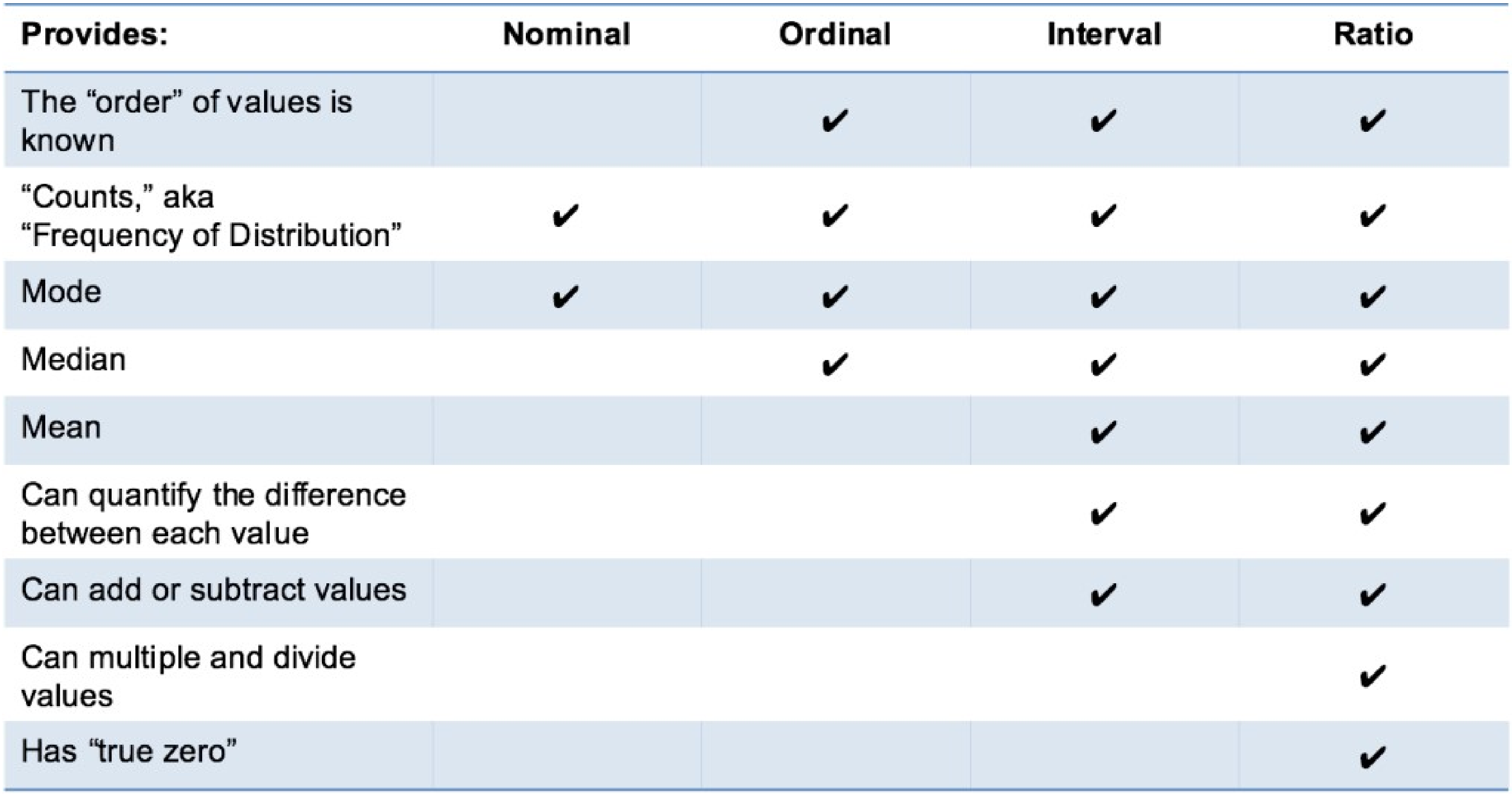
E.g. – No of Parts damaged in shipment

Continuous

- Continuous data is information that can be measured on a continuum or scale .
- Continuous data can have almost any numeric value and can be meaningfully subdivided into finer and finer increments .
- E.g. – Length, Size, width

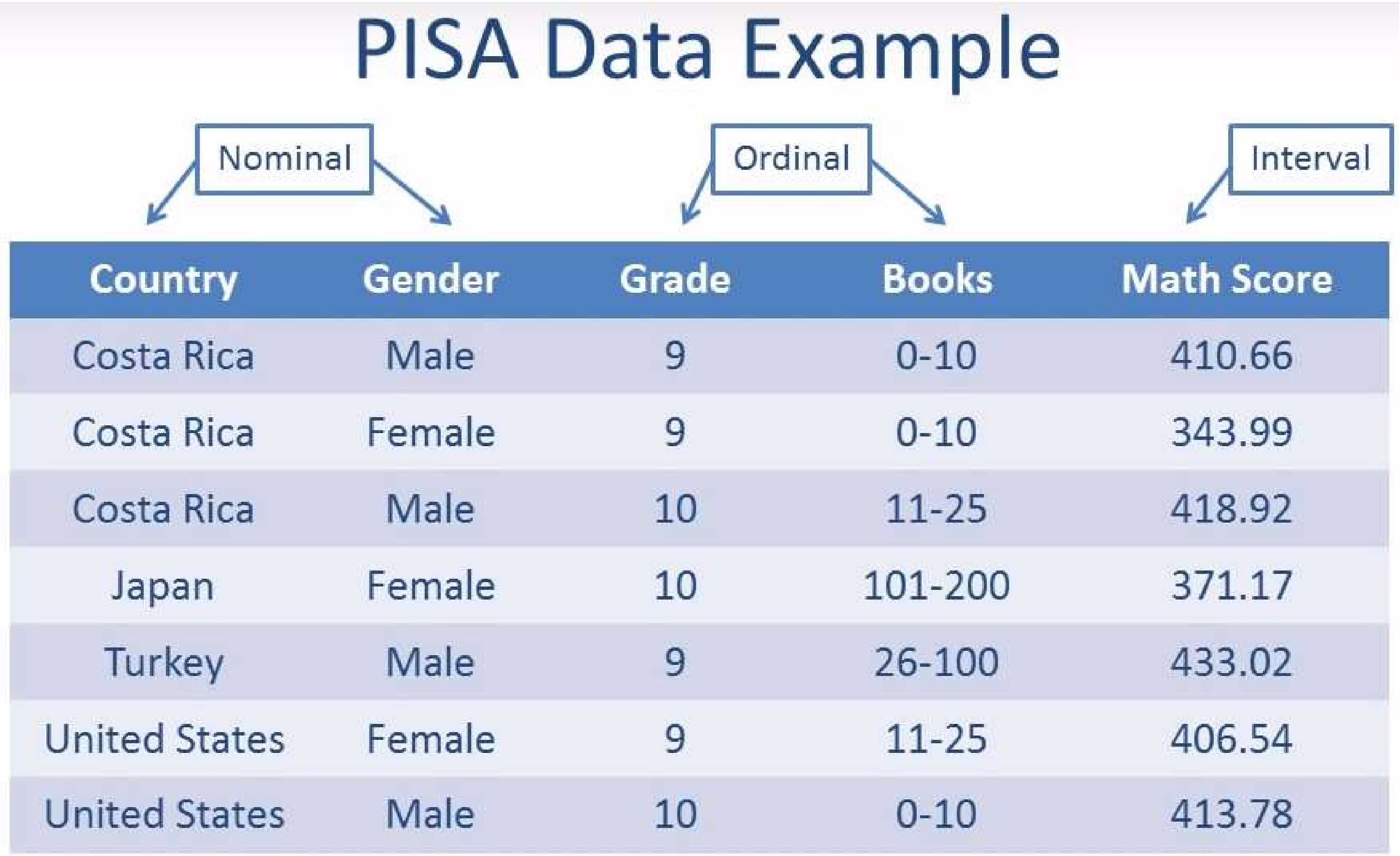


	Continuous	Discrete		
	Quantitative data	Qualitative / Categorical / Attribute data		
Measurement	**Units (example)**	**Ordinal (example)**	**Nominal (example)**	**Binary (example)**
Time of day	Hours, minutes, seconds	1, 2, 3, etc.	N/A	a.m./p.m.
Date	Month, date, year	Jan., Feb., Mar., etc.	N/A	Before / After
Cycle time	Hours, minutes, seconds, month, date, year	10, 20, 30, etc.	N/A	Before / After
Speed	Miles per hour/centimeters per second	10, 20, 30, etc.	N/A	Fast / Slow
Brightness	Lumens	Light, medium, dark	N/A	On / Off
Temperature	Degrees C or F	10, 20, 30, etc.	N/A	Hot / Cold
<Count data>	Number of things	10, 20, 30, etc.	N/A	Large / Small
Test scores	Percent, number correct	F, D, C, B, A	N/A	Pass / Fail
Defects	N/A	Number of cracks	N/A	Good / Bad
Defects	N/A	N/A	Oversized, missing	Good / Bad
Color	N/A	N/A	Red, blue, green	N/A
Location	N/A	N/A	East, West, South	Domestic / International
Groups	N/A	N/A	HR, Legal, IT	Exempt / Non-exempt
Anything	Percent	10, 20, 30, etc.	N/A	Above / Below



Provides:	Nominal	Ordinal	Interval	Ratio
The "order" of values is known		✓	✓	✓
"Counts," aka "Frequency of Distribution"	✓	✓	✓	✓
Mode	✓	✓	✓	✓
Median		✓	✓	✓
Mean			✓	✓
Can quantify the difference between each value			✓	✓
Can add or subtract values			✓	✓
Can multiple and divide values				✓
Has "true zero"				✓

Dataset Example



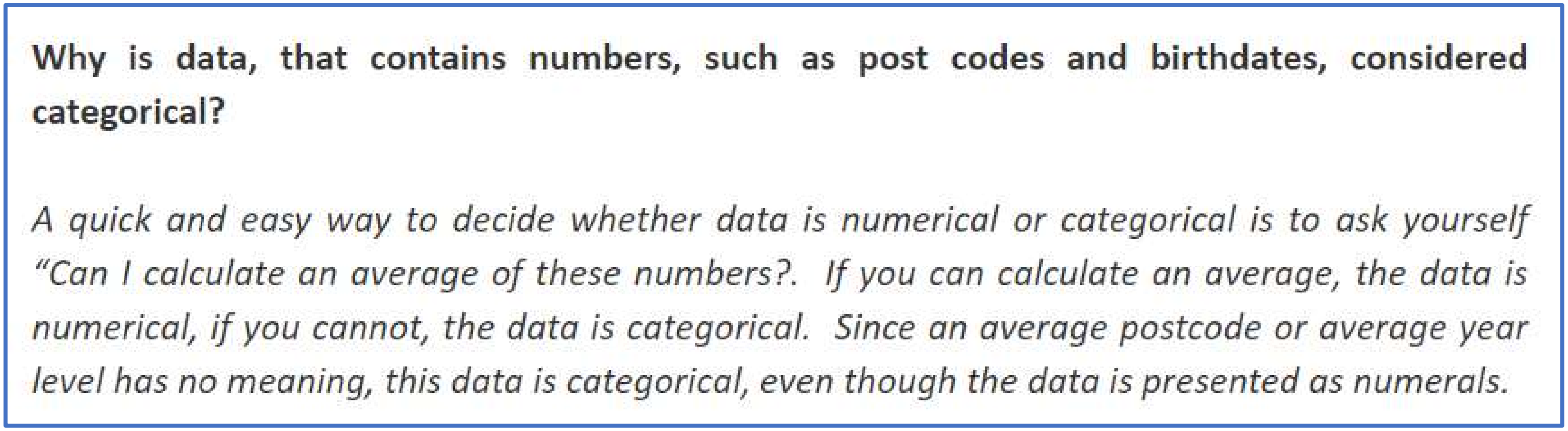
PISA Data Example

Nominal

Ordinal

Interval

Country	Gender	Grade	Books	Math Score
Costa Rica	Male	9	0-10	410.66
Costa Rica	Female	9	0-10	343.99
Costa Rica	Male	10	11-25	418.92
Japan	Female	10	101-200	371.17
Turkey	Male	9	26-100	433.02
United States	Female	9	11-25	406.54
United States	Male	10	0-10	413.78

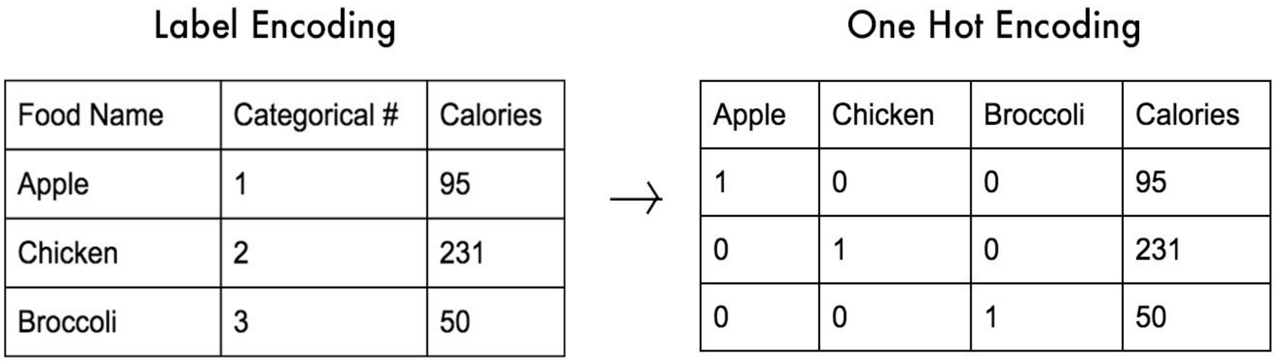


Why is data, that contains numbers, such as post codes and birthdates, considered categorical?

A quick and easy way to decide whether data is numerical or categorical is to ask yourself "Can I calculate an average of these numbers?. If you can calculate an average, the data is numerical, if you cannot, the data is categorical. Since an average postcode or average year level has no meaning, this data is categorical, even though the data is presented as numerals.

Digitalization and Feature Normalization

- Digitalization : Label Encoding, One Hot Encoding, Hash Encoding



Label Encoding

Food Name	Categorical #	Calories
Apple	1	95
Chicken	2	231
Broccoli	3	50

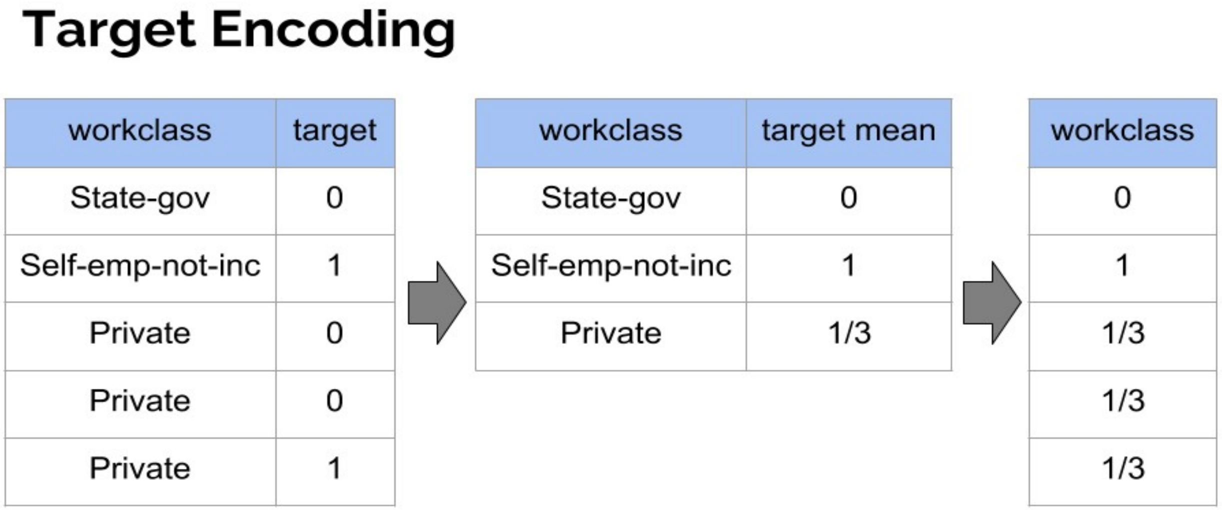
→

One Hot Encoding

Apple	Chicken	Broccoli	Calories
1	0	0	95
0	1	0	231
0	0	1	50

Digitalization and Feature Normalization

- Digitalization : Label Encoding, One Hot Encoding, Hash Encoding



Target Encoding

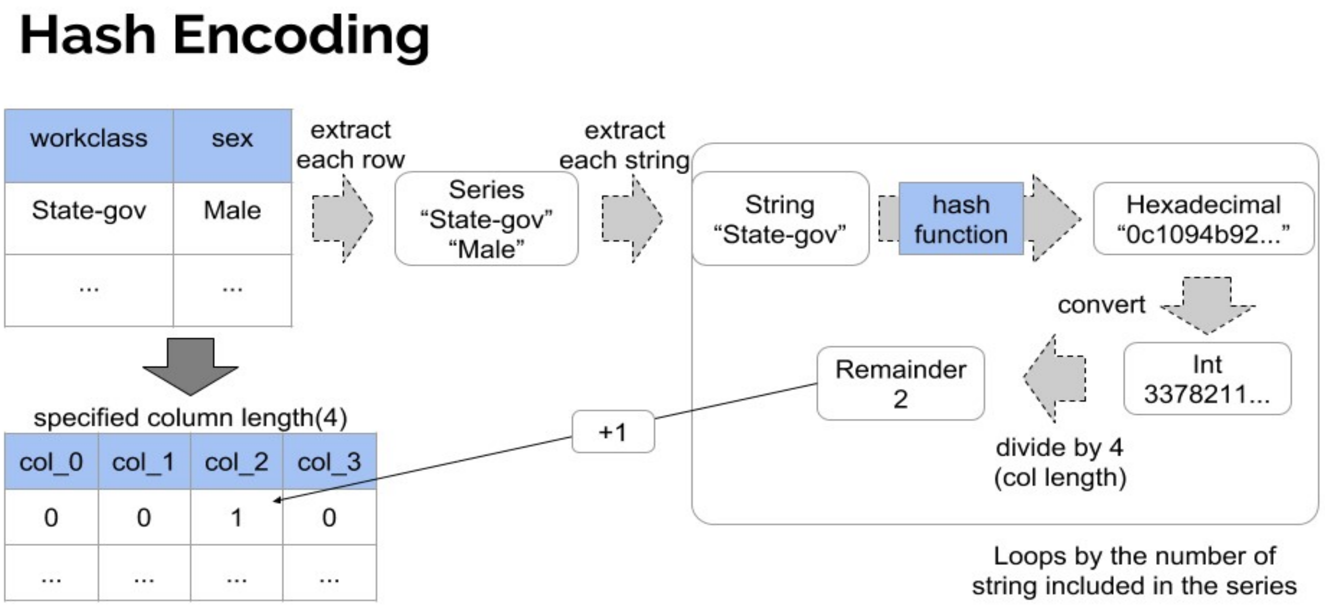
workclass	target
State-gov	0
Self-emp-not-inc	1
Private	0
Private	0
Private	1

workclass	target mean
State-gov	0
Self-emp-not-inc	1
Private	1/3

workclass
0
1
1/3
1/3
1/3

Digitalization and Feature Normalization

- Digitalization : Label Encoding, One Hot Encoding, Hash Encoding



Hash Encoding

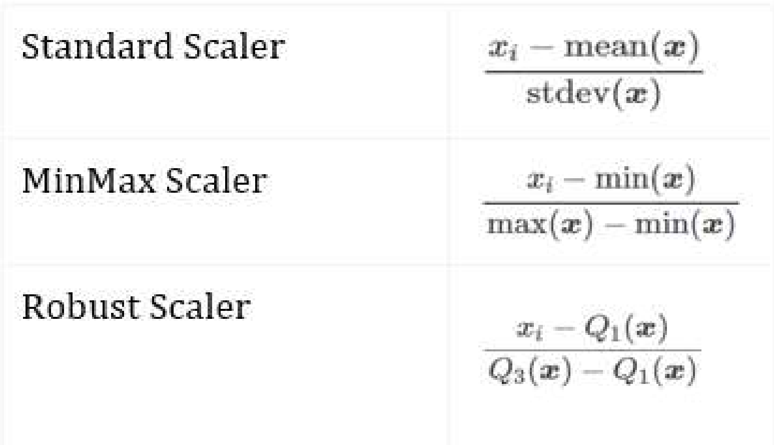
workclass	sex
State-gov	Male
...	...

specified column length(4)

col_0	col_1	col_2	col_3
0	0	1	0
...	...	...	...

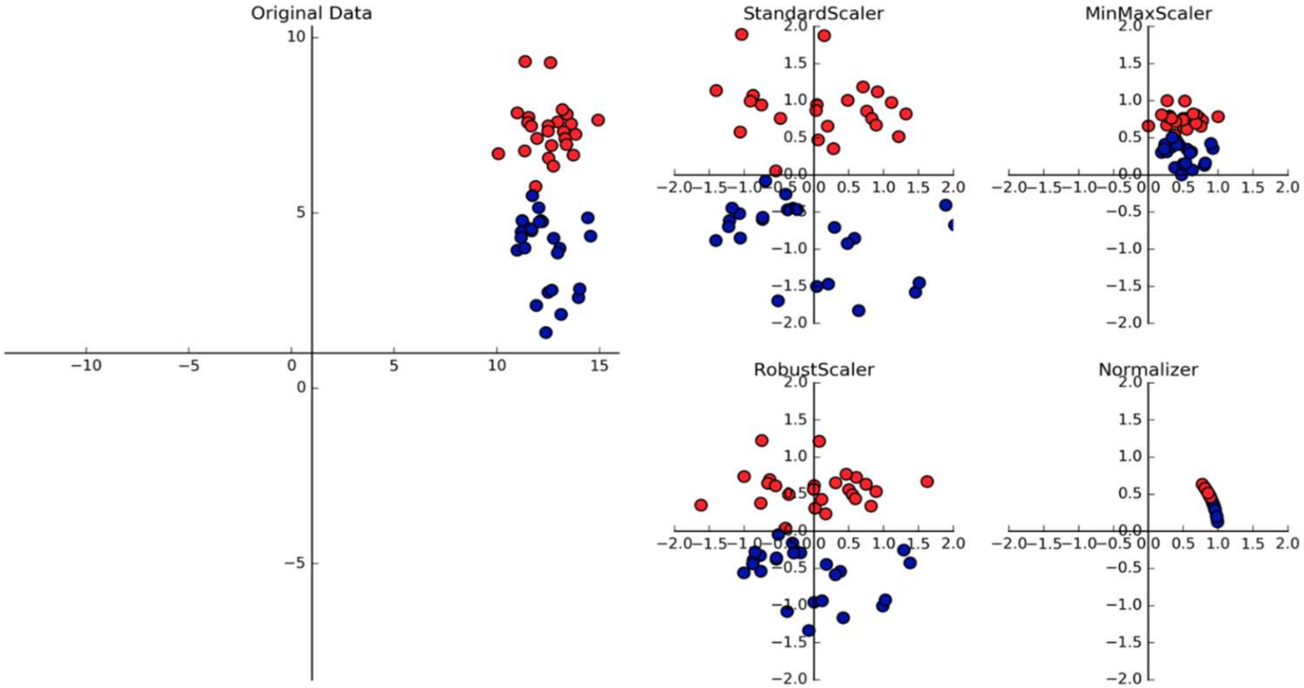
Digitalization and Feature Normalization

- **Normalization** is another important concept needed to change all features to the same scale. This allows for faster convergence on learning, and more uniform influence for all weights
  - **Standard Scaler**: This changes the data to have means of 0 and standard error of 1.
  - **Min Max Scale** : Another way to normalize is to use the Min Max Scaler, which changes all features to be between 0 and 1,
  - **RobustScaler**: Works similarly to standard scaler except that it uses median and quartiles, instead of mean and variance. Good as it ignores data points that are outliers.



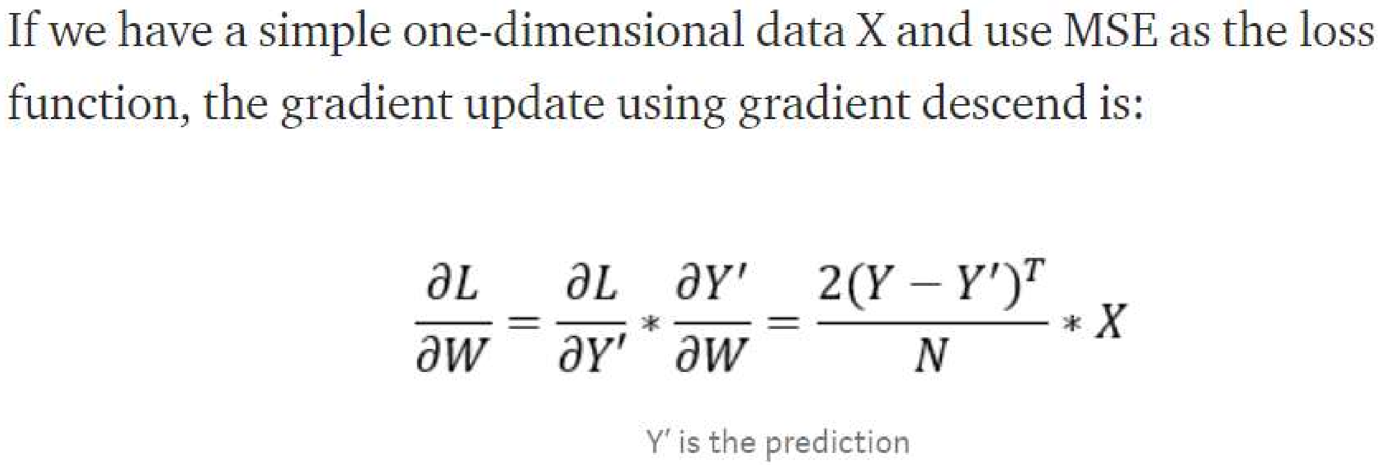
Standard Scaler	$\frac{x_i - \text{mean}(\boldsymbol{x})}{\text{stdev}(\boldsymbol{x})}$
MinMax Scaler	$\frac{x_i - \min(\boldsymbol{x})}{\max(\boldsymbol{x}) - \min(\boldsymbol{x})}$
Robust Scaler	$\frac{x_i - Q_1(\boldsymbol{x})}{Q_3(\boldsymbol{x}) - Q_1(\boldsymbol{x})}$

Digitalization and Feature Normalization



Why Normlization

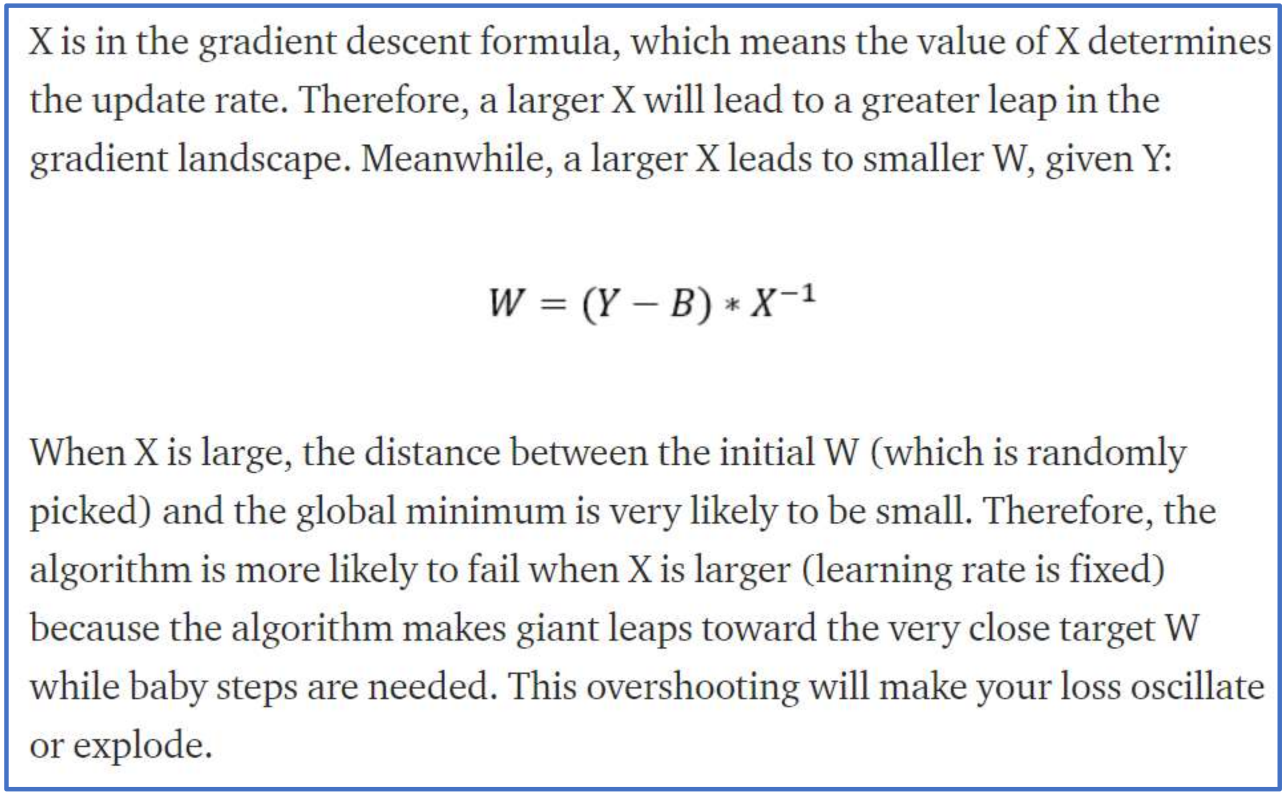
- **Standardization improves the numerical stability of your model**



If we have a simple one-dimensional data X and use MSE as the loss function, the gradient update using gradient descend is:

$$\frac{\partial L}{\partial W} = \frac{\partial L}{\partial Y'} * \frac{\partial Y'}{\partial W} = \frac{2(Y - Y')^T}{N} * X$$

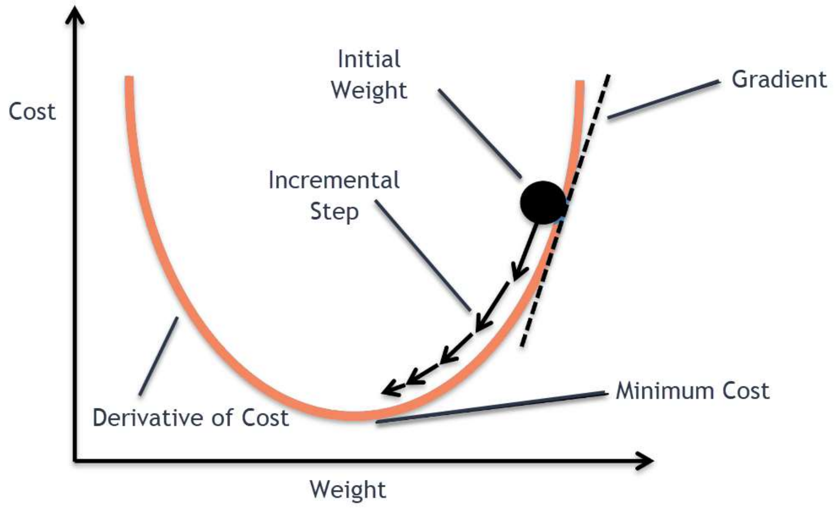
Y′ is the prediction



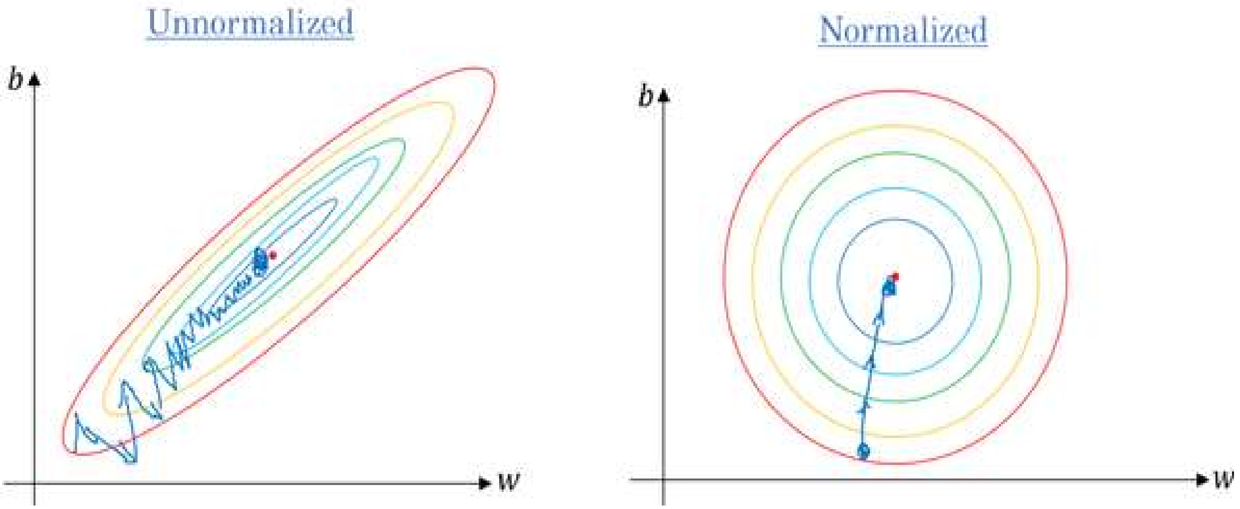
X is in the gradient descent formula, which means the value of X determines the update rate. Therefore, a larger X will lead to a greater leap in the gradient landscape. Meanwhile, a larger X leads to smaller W, given Y:

$$W = (Y - B) * X^{-1}$$

When X is large, the distance between the initial W (which is randomly picked) and the global minimum is very likely to be small. Therefore, the algorithm is more likely to fail when X is larger (learning rate is fixed) because the algorithm makes giant leaps toward the very close target W while baby steps are needed. This overshooting will make your loss oscillate or explode.

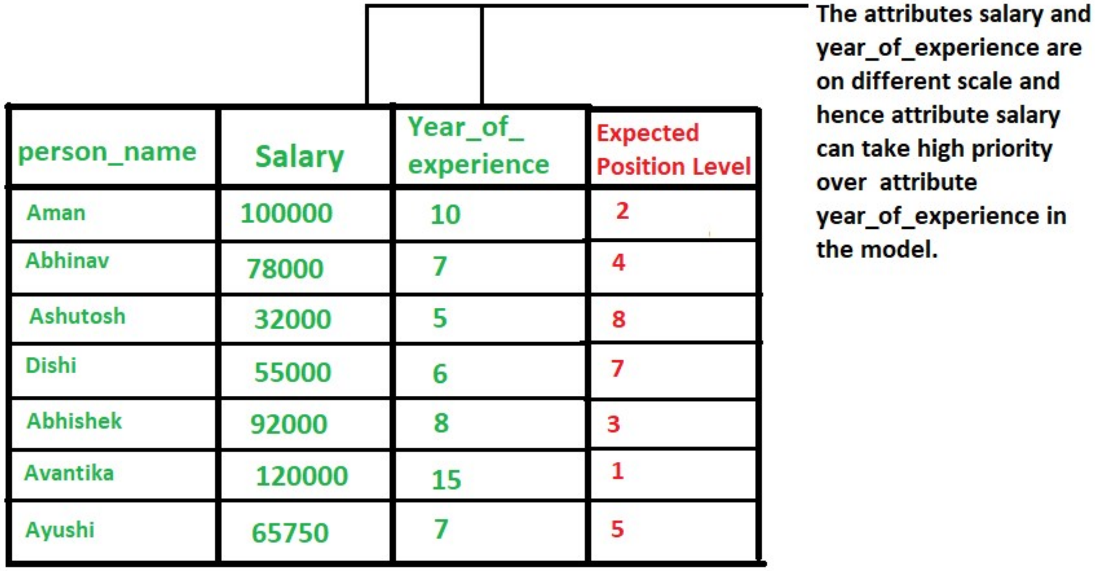


Feature scaling speeds up gradient descent by making the geometry of the cost function more hospitable ((i.e. more circular). Without feature scaling the coutour plots may not appear to be thin. In such a plot the gradient descent algorithm would oscillate a lot back and forth, taking a long time before finding its way to the minimum point.



Digitalization and Feature Normalization

- In simple words, when multiple attributes are there but attributes have values on different scales, this may lead to poor data models while performing data mining operations. So they are normalized to bring all the attributes on the same scale.

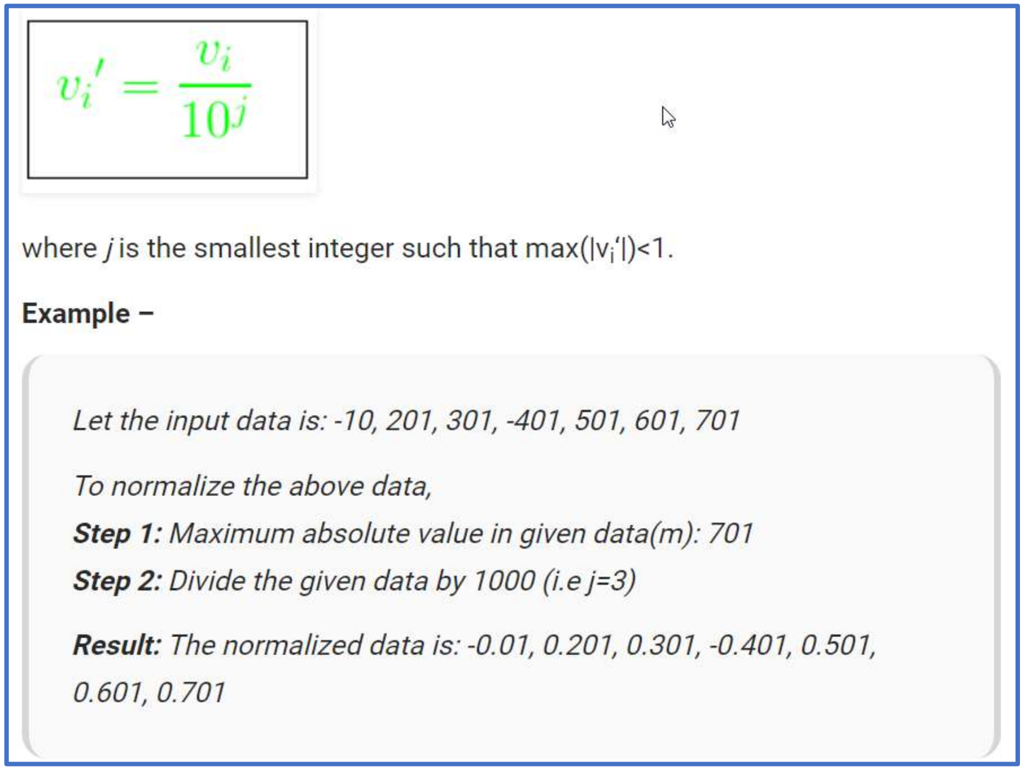


person_name	Salary	Year_of_ experience	Expected Position Level
Aman	100000	10	2
Abhinav	78000	7	4
Ashutosh	32000	5	8
Dishi	55000	6	7
Abhishek	92000	8	3
Avantika	120000	15	1
Ayushi	65750	7	5

The attributes salary and year_of_experience are on different scale and hence attribute salary can take high priority over attribute year_of_experience in the model.

Decimal Scaling Method For Normalization

- It normalizes by moving the decimal point of values of the data. To normalize the data by this technique, we divide each value of the data by the maximum absolute value of data. The data value, v_i, of data is normalized to v_i' by using the formula below –



$$v_i' = \frac{v_i}{10^j}$$

where *j* is the smallest integer such that $\max(|v_i'|)<1$.

Example –

Let the input data is: -10, 201, 301, -401, 501, 601, 701

To normalize the above data,

Step 1: *Maximum absolute value in given data(m): 701*

Step 2: *Divide the given data by 1000 (i.e j=3)*

Result: *The normalized data is: -0.01, 0.201, 0.301, -0.401, 0.501, 0.601, 0.701*

Data Reduction Strategies

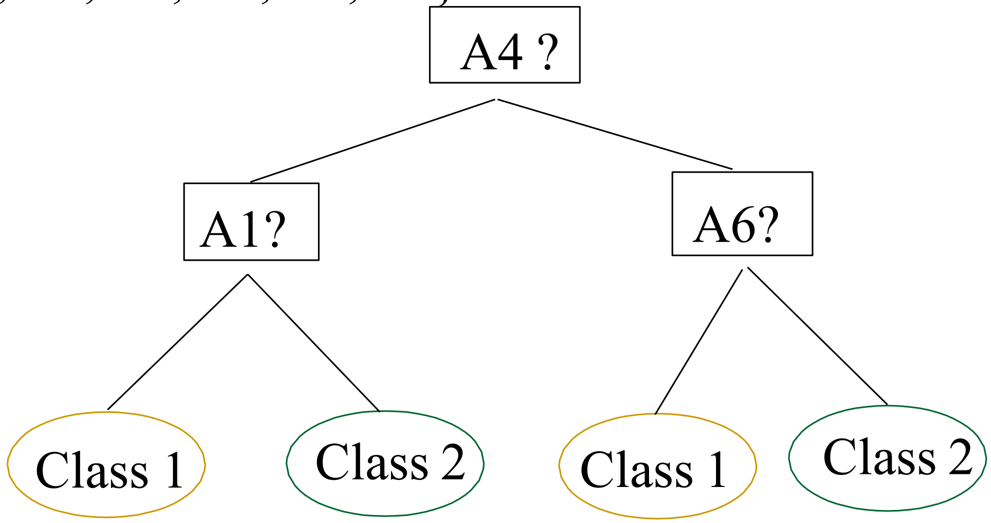
- Need for data reduction
  - A database/data warehouse may store terabytes of data
  - Complex data analysis/mining may take a very long time to run on the complete data set
- Data reduction
  - Obtain a reduced representation of the data set that is much smaller in volume but yet produce the same (or almost the same) analytical results
- Data reduction strategies
  - Data cube aggregation
  - Attribute Subset Selection
  - Numerosity reduction — e.g., fit data into models
  - Dimensionality reduction - Data Compression
  - Discretization and concept hierarchy generation

2. Attribute Subset Selection

- Feature selection (i.e., attribute subset selection):
  - Select a minimum set of features such that the probability distribution of different classes given the values for those features is as close as possible to the original distribution given the values of all features
  - reduce # of patterns - easier to understand
- Heuristic methods (due to exponential # of choices):
  - Step-wise forward selection
  - Step-wise backward elimination
  - Combining forward selection and backward elimination
  - Decision-tree induction

Example of Decision Tree Induction

Initial attribute set:
{A1, A2, A3, A4, A5, A6}



-------> Reduced attribute set: {A1, A4, A6}

Heuristic Feature Selection Methods

- There are 2^d possible sub-features of *d* features
- Several heuristic feature selection methods:
  - Best single features under the feature independence assumption: choose by significance tests
  - Best step-wise feature selection
    - The best single-feature is picked first
    - Then next best feature condition to the first, ...
  - Step-wise feature elimination
    - Repeatedly eliminate the worst feature
  - Best combined feature selection and elimination

3. Numerosity Reduction

- Reduce data volume by choosing alternative, smaller forms of data representation
- Parametric methods
  - Assume the data fits some model, estimate model parameters, store only the parameters, and discard the data (except possible outliers)
    - Example: Log-linear models, Regression
- Non-parametric methods
  - Do not assume models
    - Major families: histograms, clustering, sampling

Regression and Log-Linear Models

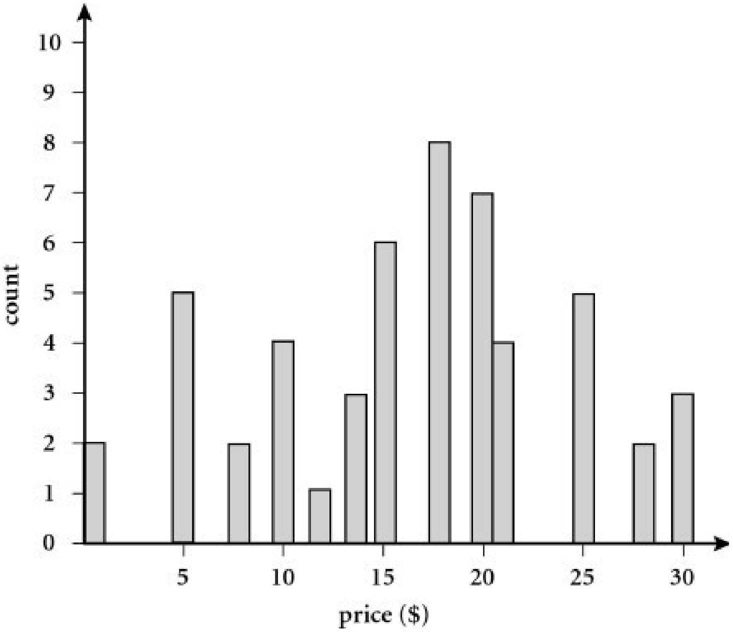
- Regression
  - Handles skewed data
  - Computationally Intensive
- Log linear models
  - Scalability
  - Estimate the probability of each point in a multi-dimensional space based on a smaller subset
  - Can construct higher dimensional spaces from lower dimensional ones
- Both
  - Sparse data

Histograms

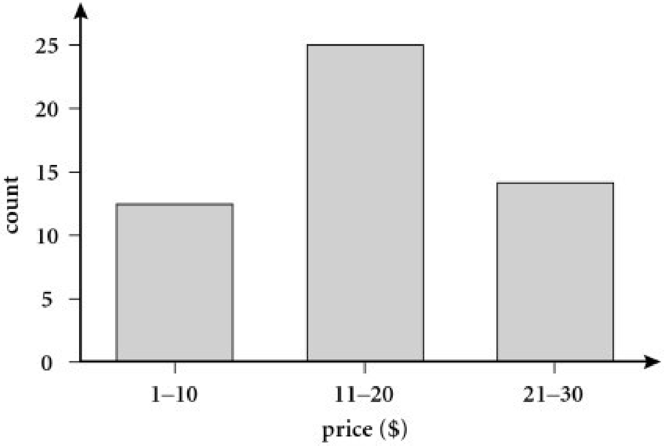
- Divide data into buckets and store average (sum) for each bucket
- Partitioning rules:
  - Equal-width: equal bucket range
  - Equal-frequency (or equal-depth)
  - V-optimal: with the least *histogram variance* (weighted sum of the original values that each bucket represents)
  - MaxDiff: Consider difference between pair of adjacent values. Set bucket boundary between each pair for pairs having the β (No. of buckets)–1 largest differences
- Multi-dimensional histogram

Histograms

Histogram with Singleton buckets



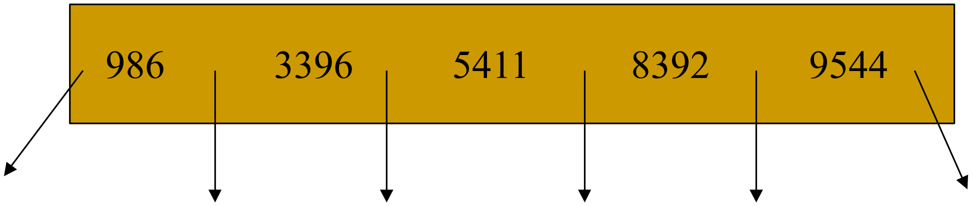
Equal-width Histogram



List of prices: 1, 1, 5, 5, 5, 5, 5, 8, 8, 10, 10, 10, 10, 12, 14, 14, 14, 15, 15, 15, 15, 15, 15, 18, 18, 18, 18, 18, 18, 18, 18, 20, 20, 20, 20, 20, 20, 20, 21, 21, 21, 21, 25, 25, 25, 25, 25, 28, 28, 30, 30, 30

Clustering

- Partition data set into clusters based on similarity, and store cluster representation (e.g., centroid and diameter) only
- Can be very effective if data is clustered but not if data is "smeared"
- Can have hierarchical clustering and be stored in multi-dimensional index tree structures
  - Index tree / B+ tree – Hierarchical Histogram



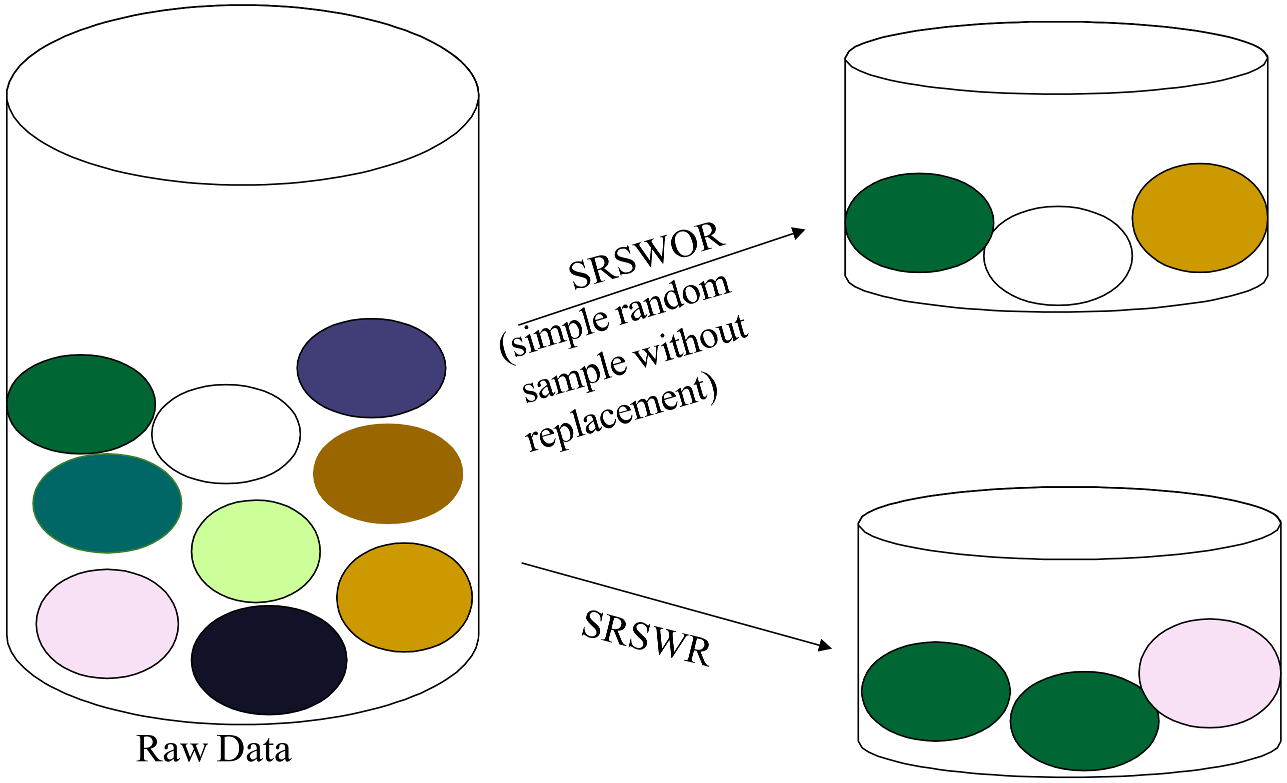
Sampling

- Sampling: obtaining a small sample *s* to represent the whole data set *N*
- Allow a mining algorithm to run in complexity that is potentially sub-linear to the size of the data
- Choose a representative subset of the data
  - Simple random sampling may have very poor performance in the presence of skew
- Develop adaptive sampling methods
  - Stratified sampling

Sampling Techniques

- Simple Random Sample Without Replacement (SRSWOR)
- Simple Random Sample With Replacement (SRSWR)
- Cluster Sample
- Stratified Sample

Sampling: with or without Replacement



Cluster Sample

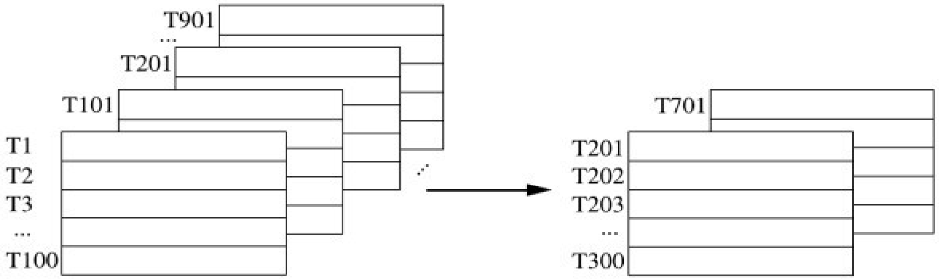
- Tuples are grouped into M mutually disjoint clusters
- SRS of m clusters is taken where m < M
- Tuples in a database retrieved in pages
  - Page - Cluster
  - SRSWOR to pages

Stratified Sample

- Data is divided into mutually disjoint parts called strata
- SRS at each stratum
- Representative samples ensured even in the presence of skewed data

Cluster and Stratified Sampling

Cluster sample
(s = 2)



Stratified sample
(according to *age*)

T38	youth
T256	youth
T307	youth
T391	youth
T96	middle_aged
T117	middle_aged
T138	middle_aged
T263	middle_aged
T290	middle_aged
T308	middle_aged
T326	middle_aged
T387	middle_aged
T69	senior
T284	senior

T38	youth
T391	youth
T117	middle_aged
T138	middle_aged
T290	middle_aged
T326	middle_aged
T69	senior

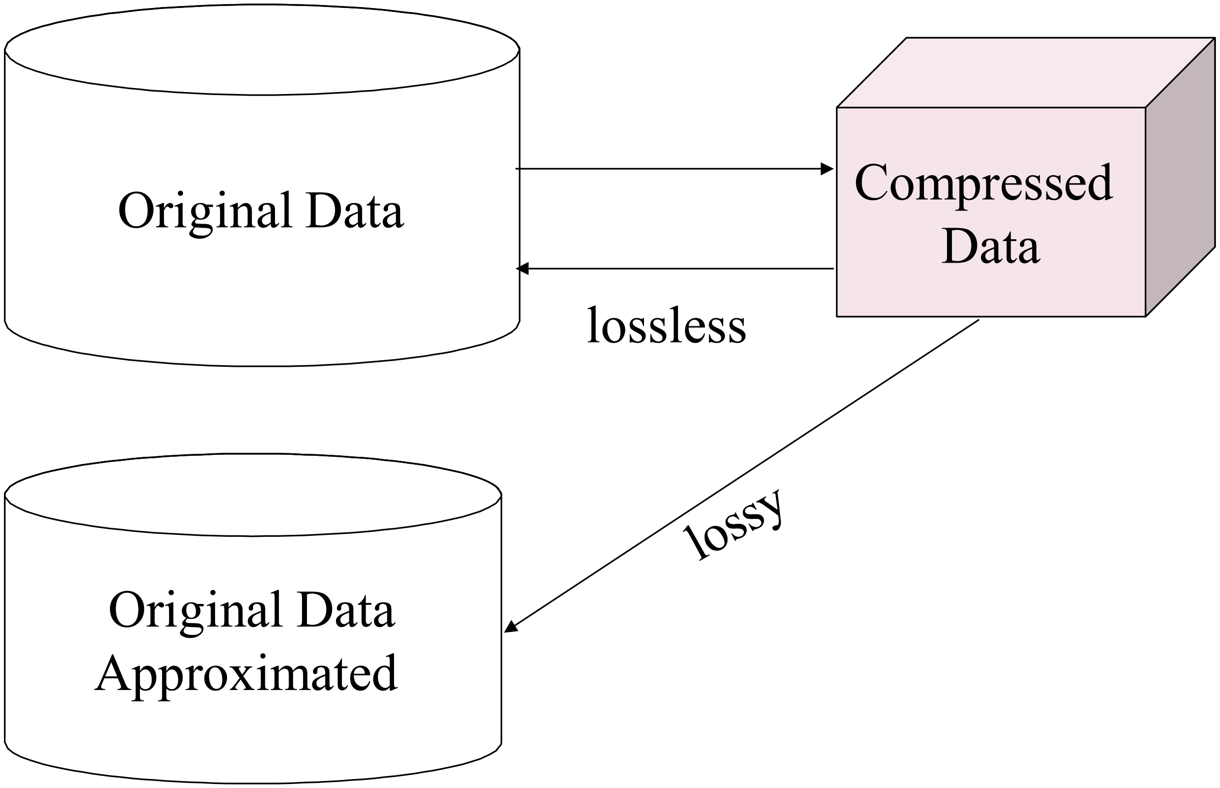
Features

- Cost depends on size of sample
- Sub-linear on size of data
- Linear with respect to dimensions
- Estimates answer to an aggregate query

Data Compression

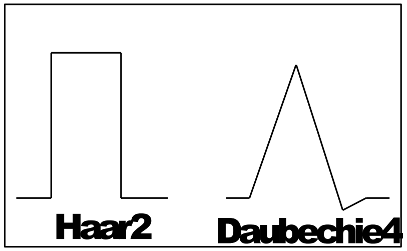
- String compression
  - There are extensive theories and well-tuned algorithms
  - Typically lossless
  - But only limited manipulation is possible
- Audio/video compression
  - Typically lossy compression, with progressive refinement
  - Sometimes small fragments of signal can be reconstructed without reconstructing the whole

Data Compression



Dimensionality Reduction: Wavelet Transformation

- Discrete wavelet transform (DWT): linear signal processing
- Compressed approximation: store only a small fraction of the strongest of the wavelet coefficients
- Similar to discrete Fourier transform (DFT), but better lossy compression, localized in space
- Method:
  - Length, L, must be an integer power of 2 (padding with 0's, when necessary)
  - Each transform has 2 functions: smoothing, difference
  - Applies to pairs of data, resulting in two set of data of length L/2
  - Applies two functions recursively



Discrete Wavelet Transform

- Can also apply Matrix Multiplication
- Fast DWT algorithm
- Can be applied to Multi-Dimensional data such as Data Cubes
- Wavelet transforms work well on sparse or skewed data
- Used in Computer Vision, Compression of finger print images

Dimensionality Reduction: Principal Component Analysis (PCA)

Given *N* data vectors from *k*-dimensions, find $c \leq$ k orthogonal vectors (*Principal components*) that can be best used to represent data

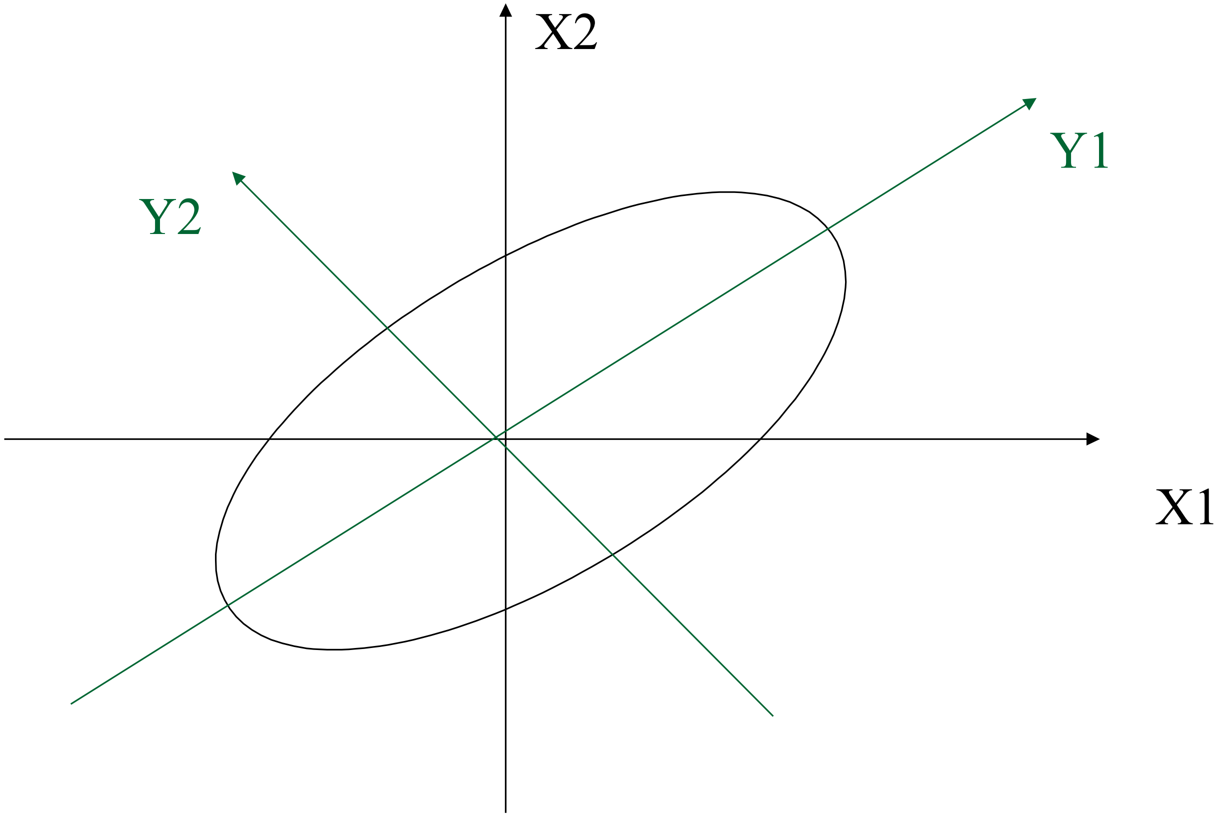
Steps

- Normalize input data: Each attribute falls within the same range

Compute *c* orthonormal (unit) vectors, i.e., *principal components*

- Each input data (vector) is a linear combination of the *c* principal component vectors
- The principal components are sorted in order of decreasing “significance” or strength
- Since the components are sorted, the size of the data can be reduced by eliminating the weak components, i.e., those with low variance. (i.e., using the strongest principal components, it is possible to reconstruct a good approximation of the original data)

Principal Component Analysis



THANK YOU